CLEAR CAUSAL EVIDENCE GUIDELINES, VERSION 1.1

The U.S. Department of Labor (DOL) established the Clearinghouse for Labor Evaluation and Research (CLEAR) to provide practitioners, policymakers, researchers, the media, and the general public with a central and trusted source of evidence from research and evaluations, including what strategies work, across a range of labor-related issues. For causal research—defined as research that attempts to assess the effectiveness of a program, policy or activity (hereafter referred to generically as an *intervention*), CLEAR provides an objective assessment and rating of the degree to which the research establishes the causal impact of the intervention. This document provides the CLEAR evidence criteria for rating the quality of causal research. A companion addendum provides additional examples of how to apply these guidelines.[1]

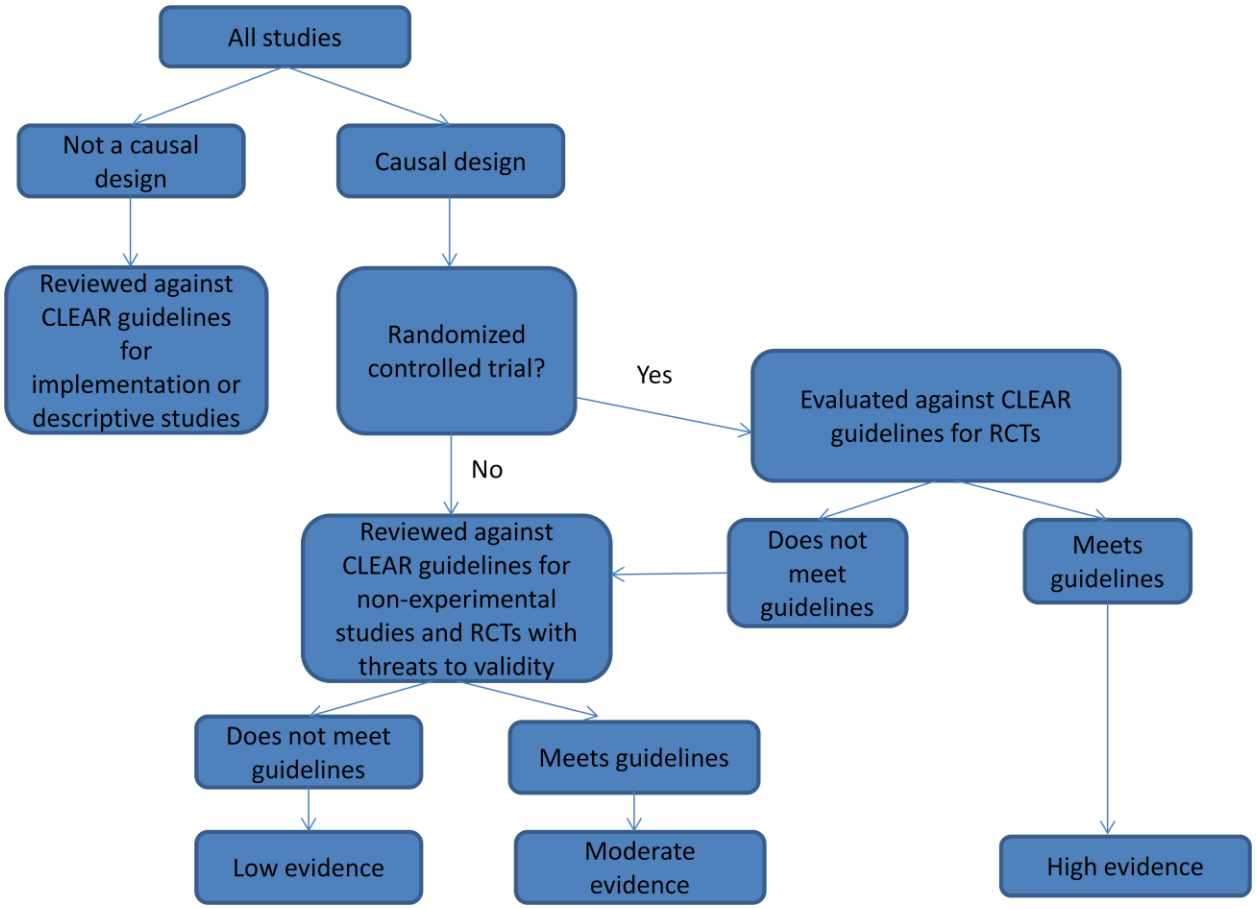
In collaboration with DOL and a Technical Work Group (TWG) of experts, Mathematica Policy Research developed a set of causal evidence guidelines to be used in reviewing nonexperimental research with causal designs. These causal designs include instrumental variables and various other regression analyses, including those with fixed or random effects and difference-in-differences. In addition to nonexperimental designs, CLEAR assesses the quality of evidence for randomized controlled trials (RCTs) using an adaptation of the Institute for Education Science's What Works Clearinghouse (WWC) standards.[2] During CLEAR's pilot phase, the evidence guidelines underwent a continuous review and improvement process and were revised to reflect lessons learned as they were first implemented. Version 1.1 incorporates these revisions, as well as feedback from DOL and the TWG for CLEAR. It also incorporates additional examples of how to apply the guidelines, gleaned from reviews in the pilot phase.

CLEAR has three possible ratings to describe the strength of causal evidence presented in a given piece of research. Well-conducted RCTs that are determined to have low attrition and no other threats to study validity receive the highest evidence rating CLEAR offers: *high*. This rating means we are confident that the estimated effects are solely attributable to the intervention that was examined. RCTs that have high attrition and/or some other threat to study validity can be evaluated against CLEAR evidence guidelines for nonexperimental designs. Research designs that meet these guidelines receive a *moderate* rating; this rating indicates that there is evidence that the study establishes a causal relationship between the intervention being examined and the outcomes of interest, but there may be other factors that were not included in the analysis that also could affect the outcomes of interest. Research that does not meet the criteria for a high or moderate rating receives a *low* rating, which indicates that we cannot be confident that the estimated effects are attributable to the intervention being examined. Figure 1 summarizes how different causal designs are reviewed and the evidence ratings they are eligible to receive.

[1] The guidelines presented here are for comprehensive, second-level reviews, as described in the CLEAR Policies and Procedures. First-level reviews are used to produce brief highlights of research that do not include a rating of the quality of the causal evidence and thus do not require use of these guidelines.

[2] In the future, CLEAR guidelines will also include an adaptation of WWC pilot standards for regression discontinuity designs.

Figure 1: CLEAR Causal Evidence Rating Scheme



A. Considerations for All Second-Level Reviews of Causal Studies

The criteria in these guidelines are used to review evidence from research papers and reports that span a broad range of social science disciplines over many years. For this reason, with few exceptions, the guidelines generally do not require specific approaches, such as particular specification tests or standard error corrections. Rather, they provide general criteria that must be met, and a supporting addendum provides examples illustrating how the criteria could be satisfied.

In many cases, determining whether a given criterion has been met requires judgment on the part of a reviewer, because there is no definitive test to indicate that it has been met. In such cases, the guidelines specify that the study must make a convincing case that the criterion is satisfied; if reviewers can identify a plausible scenario under which the criterion is not satisfied, the study fails this test. In these cases, reviewers note their concerns and the Principal Investigator (PI) for the topic area makes a judgment as to whether the concerns are substantial enough to threaten the causal validity of the study. In making this determination, the PI considers the likelihood that other factors affected the outcome and the potential magnitude of those effects.

In addition to the criteria for meeting the threshold of causal validity, there are other considerations around calculation of standard errors that the reviewers take into account. For

example, in the case of instrumental variables analyses, the reviewers consider whether standard errors were calculated using an appropriate method, such as the delta method or bootstrapping, and whether, if necessary, standard errors accounted for the first stage of estimation in a two-step process. For analyses using longitudinal data, the reviewers consider whether the authors calculated standard errors using a method that accounts for serial correlation, heteroskedasticity and different levels of aggregation (for example, cluster-robust standard errors for designs that use individual and state data). However, these considerations do not affect the causal evidence rating. If the report authors did not report standard errors that were computed in an appropriate manner, the summary produced by CLEAR notes this as a concern.

Finally, these causal evidence ratings relate only to the extent to which a given study demonstrates a causal effect ("internal validity") and not the extent to which that causal effect would be expected in different contexts ("external validity"). The summaries produced as a result of CLEAR reviews of causal research designs provide information about the specific context in which the study was conducted so that readers may draw their own conclusions about whether the findings would apply to other situations. Similarly, issues with data quality, such as low response rates on a survey, would be discussed in the summary but would not affect the causal evidence rating (unless they caused the study to not meet one or more evidence criteria).

The sections that follow describe the guidelines for instrumental variables and a broad set of other regression methods including fixed effects models, random effects models, difference-in-differences models, and matched comparison group designs. The final section describes how the WWC evidence standards were adapted for use in CLEAR ratings of RCTs. An addendum to the guidelines provides additional examples and identifies specific approaches that may be used to address the criteria described in the guidelines.

B. Instrumental Variables Models

Instrumental variables techniques are often used when receiving an intervention—that is, participating in a program or being subject to a policy—is determined by a combination of endogenous and exogenous factors. Endogenous factors are things that are related to both receiving the intervention *and* the outcomes of interest in a study. For example, highly motivated individuals may be more likely to participate in a job training program but also tend to have better outcomes even without the intervention. Exogenous factors are things that are related to receiving the intervention but are *not* related to outcomes. For example, an exogenous factor could be a lottery to select which program applicants will be admitted to an over-subscribed program.

If receiving the intervention is related to unmeasured characteristics or conditions that influence outcomes, a simple regression of outcomes on an indicator of receiving the intervention will lead to biased estimates of the causal relationship. For example, individuals may self-select for a training program or firms may be selected for monitoring by an enforcement agency by a nonrandom process. A simple analysis that does not account for this relationship will inappropriately attribute all observed effects to the program. One approach to dealing with this problem is to estimate impacts using only exogenous factors that affect whether the individual received the intervention while filtering out the influence of the endogenous factors. These exogenous factors are sometimes called "instrumental variables." Multiple methods exist for estimating impacts using instrumental variables

including two stage least squares, the Heckman two-step correction, and limited information maximum likelihood.

Criterion IV.1: Sufficient instrument strength (or, instrument relevance). The instrument must have sufficient strength to predict whether the intervention was received or impact estimates may be biased. Therefore, to meet this criterion studies must report a test of instrument strength. One common test uses a first-stage equation that models treatment as a function of the instrument and all explanatory variables. The test is based on the first-stage F-statistic for the null hypothesis that the instrument has no effect on treatment. If the F-statistic exceeds 10, the instrument is considered to be of sufficient strength. See the addendum for other ways to test instrument strength.

Criterion IV.2: Exclusion restriction (or, instrument exogeneity). The only plausible means by which the instrument affects the outcome must be through receipt of the intervention. The study must make a clear and convincing case that the exclusion restriction is satisfied; otherwise the reviewer assumes that it is not. If the reviewer identifies a plausible mechanism by which the instrument would influence the outcome directly (after controlling for observed factors), the research design does not satisfy this criterion.

In addition, research designs with multiple endogenous variables and instruments must satisfy a third criterion:

Criterion IV.3: Identification. There must be at least as many instruments as there are endogenous explanatory variables, and the instruments must not be highly collinear with each other. In technical terms, this is often described as "satisfying the rank condition." See the addendum for examples of how to test whether this condition is satisfied.

C. Regression Analyses

Regression analyses attempt to estimate causal effects by controlling for enough explanatory variables that the estimated treatment effect is plausibly uncorrelated with the error term. Regression analyses can use several different techniques, including ordinary least squares, probit, logit, tobit, matching methods, and hazard models. In addition to these methods, the guidelines offer extra guidance for special cases of regression techniques including fixed effects, difference-in-differences, and random effects models.

Fixed effects (FE) are used to take into account unobserved, time-invariant characteristics of sample members that might affect both whether they received the intervention—that is, whether they participated in a program or were subject to a policy—and the outcomes of interest. For example, choosing to participate in a job training program could be a function of an individual's unobserved motivation and cognitive ability, which could also be related to the individual's earnings-related outcomes; the results of an analysis that did not take this into consideration would be biased. In this example, including an individual fixed effect in the model would account for all time-invariant characteristics (such as motivation) of the individuals.

Other types of fixed effects models examine an intervention that was applied at the group level, such as a state or a firm, and where individuals may be affected by the intervention by virtue of being in the entity that was subject to it without having directly opted to participate in it. In these models, fixed effects at the group level take into account unobserved, time-invariant characteristics

of the groups that might affect both receiving the intervention and the outcomes of interest. For instance, consider an analysis of the effect of minimum wage laws on the earnings of workers in the state. States can choose whether to adopt the law, but the individuals who work in that state are subject to the law. An analysis of individual workers' earnings might include a state-level fixed effect that would hold constant the other factors present in the state, such as a strong union presence, that might influence both the decision to adopt the minimum wage law and workers' earnings.

A difference-in-differences (DID) model is a special case of a fixed effects model. DID estimators measure the changes in outcomes over time for the intervention group relative to the changes over the same period for a comparison group that did not receive the intervention (or received a different one).[3] DID models often assess an intervention adopted at a group level, such as the implementation of a policy at the state level, and the analysis also takes place at the group level. For example, suppose states adopted a policy to increase the minimum wage over a ten-year period, with some adopting it in each year and some never adopting it, and a study analyzed the effect of the law on the state unemployment rate. A DID model would compare the change in unemployment rates in states that adopted the minimum wage policy to the changes in the unemployment rates in states that did not adopt the policy. This approach accounts for changes in the outcome variable that would have occurred over time for reasons unrelated to the policy as well as for time-invariant differences between the intervention and comparison groups that are unrelated to receiving the intervention.[4]

A random effects (RE) model is similar to a FE model in that it models an individual-specific effect. However, RE relies on the assumption that these time-invariant unobserved characteristics are not correlated with other explanatory variables in the model.

I. Criteria for All Regression Models

The following criteria covering threats to causal validity are used to evaluate all studies that use regression models.

Criterion Regression.1: Comparability of treatment and control groups before the intervention. The treatment and control groups must be similar before the implementation of the intervention so that the experiences of the control group present a valid picture of what would have happened to the treatment group if it had not been exposed to the intervention. There are two types of comparability that are relevant for the determining causal validity: comparability on observed characteristics and comparability on unobserved characteristics.

Observed characteristics. Comparability on observed characteristics means that the two groups are similar on key characteristics, or that the study has adjusted for differences between them by including appropriate controls in the regression. To establish comparability of the groups on observed characteristics, study authors could compare characteristics measured before the intervention for the two groups and show that the differences between the two are not statistically

[3] DID models could also be described as a short interrupted time series design with a comparison group or a comparison group design with pre-intervention and post-intervention data.

[4] Designs in which individuals receive a treatment, then stop receiving it, then receive it again, present a separate set of methodological issues that are outside the scope of these guidelines at the current time.

significant and not large enough to warrant concern. If the authors do not attempt this type of demonstration, or if the groups do appear to be different, then the authors must also control for these characteristics in the analysis. Typically, basic demographic information alone will not suffice to establish comparability of the groups or as sufficient controls in a regression; pre-intervention and/or lagged values of the key outcome measure (or fixed effects that account for pre-intervention outcomes) will usually be necessary to satisfy this criterion.

The number of lags in the pre-intervention outcomes and the types of control variables required vary by the topic area and outcome being examined and are specified in each topic area review protocol. For example, the specified pre-intervention characteristics for research that analyzes the employment outcomes of youth for the Opportunities for Youth topic area include pre-intervention measures of employment (lagged employment variables), age, gender, race/ethnicity, and geographic location. In cases where pre-intervention measures of the outcomes are not available, judgment is required as to which control variables, and the appropriateness of the functional form they take, are sufficient. Reviewers and the PI will consider these situations on a case-by-case basis.

If authors estimate models including fixed effects or lags of the dependent variable in the regression, they may instead demonstrate equivalent *trends* in pre-intervention outcomes between treatment and control groups. That is, if one identifies the effects of an intervention using changes in an outcome over time, then only the changes in that outcome prior to the intervention need to be the same across treatment and control groups, and not necessarily the levels of pre-intervention outcomes. This can be done by an inspection of pre-intervention trends or, in the case of only one pre-intervention period, the use of placebo tests (see the addendum for examples). In addition, studies using panel data must adequately control for time-varying characteristics that might influence the outcomes of interest, as specified in the topic area review protocol, and time trends when appropriate.

Unobserved characteristics. Another consideration for research designs to meet this criterion is comparability of the treatment and control groups on unobserved characteristics. This guards against situations where the treatment and control groups appear to be similar on observed characteristics, but there is an obvious selection mechanism whereby individuals select to enter the treatment group based on a characteristic that is not observed and influences both the decision to participate and the outcome of interest. For example, suppose that offers to participate in a job training program are given to applicants who meet basic screening criteria in terms of education and previous work experience. Then, only half of those offered participation in the program actually complete the program, with the other half dropping out. In this instance, those who completed the program and those who dropped out might look very similar on observed characteristics. However, those who actually completed the program likely have some innate characteristic, such as motivation or ambition, that those who dropped out did not possess, and this trait would influence both participation in the program and post-intervention outcomes. Therefore, using the dropouts as a comparison group for those who completed the program does not provide a valid picture of how the program completers would have fared in the absence of the offer to participate in the program.

Typically, any intervention that would be triggered by changes in the outcome variable will likely have issues related to non-comparability of unobservable characteristics. More generally, if reviewers

can identify a plausible selection mechanism that is not controlled for in the analysis (e.g., by including multiple lagged outcomes), the study does not meet this criterion.

Criterion Regression.2: Confounding factors. With the exception of the intervention, the changes in conditions for the comparison group should be the same as those experienced by the intervention group. Therefore, observable factors (such as simultaneous interventions) that substantially influence the outcome and that may differ between the groups must be accounted for in the analysis. For instance, suppose training participants were able to continue to collect unemployment insurance payments while participating in the program while non-participants could not collect these payments. This would confound estimates of the effect of the job training program with the effect of receiving unemployment insurance payments.

Confounding factors also include time-varying factors that differentially affect the treatment group. For instance, in the case of state policy variation, there should not be differences in other state policies occurring over the same time period that also affect the outcome of interest. If the reviewer identifies a plausible confounding factor, the research design does not meet this criterion.

Criterion Regression.3: Anticipating the intervention. The study must convince the reviewer that individuals or groups who received the intervention would not have behaved differently in anticipation of it (for example, by discussing why individuals or groups would be unable to anticipate the intervention or showing that individuals or groups did not behave differently in anticipation), or they must adjust for the anticipation appropriately. For example, suppose a new state safety standard was announced that would go into effect in six months. During the six months between policy adoption and enforcement of the new standard, businesses might begin increasing their safety precautions in anticipation of the new law, potentially reducing their rate of workplace injuries; an analysis that used six-month lagged data as controls in the regression model could therefore yield biased impact estimates. The study must be convincing that anticipating the intervention is not an issue or, if it is an issue, the analysis must address it effectively to meet this criterion.

II. Special Criterion for Estimates of Group-Level Effects

Research designs that include group-level fixed effects rather than individual-level fixed effects (for instance, using data from the Current Population Survey on individuals' earnings to evaluate the effect of state minimum wage laws) or group-level control variables to account for pre-intervention lags in the outcome variable (for instance, data from the Current Population Survey on average earnings at the state level prior to changes in the minimum wage law) must satisfy Criteria Regression.1 through Regression.3 and an additional criterion:

Criterion Regression.4: Changes in group composition. The composition of the treatment and control groups should not change in ways related to the outcome of interest. For example, changes in minimum wage laws may induce some workers to leave some states and enter neighboring ones; if this selective migration were substantial enough, the resulting estimates would conflate the direct effects of the law itself with changes in the composition of the labor force. In this case, the study must provide evidence that there was not substantial selective migration into or out of the states affected by the policy change. In cases where migration is shown to be within the cutoffs established by the WWC for attrition standards, the analysis does not need to make

additional adjustments. However, if migration exceeds the cutoffs, data on individual or group characteristics must be used to account for measurable changes in composition.

III. Special Criterion for Random Effects

In addition to Criteria Regression.1-Regression.3 (and Regression.4 if the analysis includes groups), RE models must satisfy an additional criterion:

Criterion RE.1: Use of RE over FE. When the RE model is valid, the FE estimator will still produce unbiased estimates of the relationships of interest, but they will be less efficiently estimated than the RE estimates. But if unobserved characteristic are correlated with explanatory variables in the model, the RE estimates will be biased. Therefore, in general, the FE model is preferred to RE unless there is compelling evidence that the author has included all the time-invariant factors that could be correlated with unobserved characteristics that affect the outcome. If the reviewer identifies a factor that could be correlated with other explanatory variables but is not included in the RE model, the study does not meet this criterion. In addition, studies using random effects must report a specification test justifying the use of RE over FE.

IV. Note on Matching Designs

Matching and weighting designs, including propensity score matching designs, are evaluated according to the guidelines for regression analyses. Matching analyses must match on all the control variables specified in the topic area protocol; those that match on all the specified control variables do not also need to include them in the regression as long as the match was successful. If the matching analysis does not attempt to match on one or more key control variables, or if the matching process was unsuccessful on one or more key control variables, the regression analysis must include them. A matching process is determined to be unsuccessful if, after matching, there remain statistically significant differences between the intervention and comparison group at the 5 percent level. In addition, matching methods must meet criteria Regression.1-4, as applicable.

In some instances, authors may choose to apply weights derived from a matching process (e.g., propensity score weights) to study data for analysis. In these cases, comparability of the groups on observed variables must be demonstrated on the weighted data, in addition to meeting the other applicable regression criteria.

D. Randomized Control Trials

CLEAR uses WWC evidence standards, adapted for use in a labor context, to evaluate the strength of causal evidence of RCTs. The WWC standards for RCTs have been extensively reviewed and tested, and they represent the current state-of-the art in rating the strength of RCTs.[5]

Criterion RCT.1: Sample Attrition. Sample attrition is a key factor in determining the strength of evidence for RCTs. CLEAR considers both the overall sample attrition rate and the differential in sample attrition rates between the treatment and control groups, as both contribute to the potential bias of the estimated effect of an intervention. There are conservative and liberal standards for acceptable levels of attrition. The conservative standards are applied in cases where there is reason to believe that relatively more of the attrition may be endogenous to the intervention examined—for example, disadvantaged youth choosing whether to participate in a residential career training program. The liberal standards are applied in cases where there is reason to believe that relatively little of the attrition is endogenous to the intervention examined—for example, employer cutbacks of availability of training program slots caused by reduced funding.

Attrition rates are based on the number of sample cases used in the analysis sample with measured, as opposed to imputed, values of the outcome measures. For a given level of overall attrition, Table 1 presents the maximum differential attrition rate between the treatment and control groups that is acceptable. The higher the rate of overall attrition, the lower the rate of differential attrition must be in order to be considered acceptable.

Studies based on cluster random assignment designs must meet attrition standards for both the study sample units that were assigned to intervention or comparison group status (e.g., schools or communities) and the study sample units for analysis (e.g., youth). In applying the attrition standards to the unit of analysis, the denominator for the attrition calculation includes only sample members in the clusters that remained in the study sample.

Criterion RCT.2: Confounding factors. If random assignment was properly implemented, then the only thing that differed between the treatment and comparison groups was the intervention itself. However, RCTs can still have confounding factors that do not allow the disentanglement of the effect of the intervention from some other factor. For example, if a school-wide program for youth were implemented in only one school, it would be impossible to separate the effect of the program from the effect of the staff and environment at that school.

[5] The full set of WWC evidence standards is documented in the WWC Procedures and Standards Handbook, which can be found at http://ies.ed.gov/ncee/wwc/documentsum.aspx?sid=19. The Handbook explains the criteria for evaluating RCTs, which mainly involves determining study attrition and any other threats to the validity of the study's design. The WWC standards have been adapted by other federal research clearinghouses: the U.S. Department of Health and Human Services (DHHS) for the Teen Pregnancy Prevention evidence reviews, IES for evaluation of Investing in Innovation (i3) evidence, and the DHHS Office of the Administration for Children and Families for the Home Visiting Evidence of Effectiveness systematic reviews.

Table 1. Thresholds of Acceptable Combinations of Overall and Differential Attrition

Overall Attrition (percent)	Differential Attrition		Overall Attrition (percent)	Differential Attrition	
	Conservative Boundary (percent)	Liberal Boundary (percent)		Conservative Boundary (percent)	Liberal Boundary (percent)
0	5.7	10.0	34	3.5	7.4
1	5.8	10.1	35	3.3	7.2
2	5.9	10.2	36	3.2	7.0
3	5.9	10.3	37	3.1	6.7
4	6.0	10.4	38	2.9	6.5
5	6.1	10.5	39	2.8	6.3
6	6.2	10.7	40	2.6	6.0
7	6.3	10.8	41	2.5	5.8
8	6.3	10.9	42	2.3	5.6
9	6.3	10.9	43	2.1	5.3
10	6.3	10.9	44	2.0	5.1
11	6.2	10.9	45	1.8	4.9
12	6.2	10.9	46	1.6	4.6
13	6.1	10.8	47	1.5	4.4
14	6.0	10.8	48	1.3	4.2
15	5.9	10.7	49	1.2	3.9
16	5.9	10.6	50	1.0	3.7
17	5.8	10.5	51	0.9	3.5
18	5.7	10.3	52	0.7	3.2
19	5.5	10.2	53	0.6	3.0
20	5.4	10.0	54	0.4	2.8
21	5.3	9.9	55	0.3	2.6
22	5.2	9.7	56	0.2	2.3
23	5.1	9.5	57	0.0	2.1
24	4.9	9.4	58	-	1.9
25	4.8	9.2	59	-	1.6
26	4.7	9.0	60	-	1.4
27	4.5	8.8	61	-	1.1
28	4.4	8.6	62	-	0.9
29	4.3	8.4	63	-	0.7
30	4.1	8.2	64	-	0.5
31	4.0	8.0	65	-	0.3
32	3.8	7.8	66	-	0.0
33	3.6	7.6	67	-	-

E. Research that Does Not Have a Causal Design

Many studies that claim to examine effectiveness use estimation methods that do not support causal inference. For example, studies with before-after designs fail to account for changes that might have occurred for reasons other than receiving the intervention.[6] In these cases, the evidence does not support causal conclusions regarding the impact of the intervention. Studies using such methods are reviewed under CLEAR Guidelines for Descriptive studies, and not Causal Evidence Guidelines.

[6] Interrupted time series designs, which are similar to before-after designs but use data covering more time periods, can provide evidence of causal effects under certain circumstances. CLEAR will develop evidence guidelines for this particular design if we encounter studies that use this methodology.